Project Report: Linear Regression - Insurance Dataset

Disclaimer

The following project report was written to satisfy the requirements of the final project for the Coursera Supervised Machine Learning: Regression course by IBM. It closely follows the materials outlined in the course including worked examples.

My code to produce figures as well as the figures are available in my github repository.

Objectives

For this exercise, I am training several linear regression models. The main purpose of this exercise is to see whether some features in a dataset (see description below) are able to **explain** the behavior of medical charges with respect to features in the dataset.

Description of Data Set

I am using the Medical Insurance data set available on Kaggle. The dataset consists of 1338 rows and 7 variable columns. The columns represent:

- Age, ordinal data with values between 18 and 64)
- Gender, categorical data consisting of 51% male and 49% female
- BMI (body mass index), numerical data between 16 and 53.1
- Number of children, ordinal data between 0 and 5
- Smoker, categorical data whether someone is a smoker
- Region, categorical data depending on the region ()
- Brief description of the data set you chose and a summary of its attributes.
- Insurance cost, numerical data of annual cost for insured person

The dataset does not contain any missing values.

Data Exploration

After having confirmed that the dataset does not have missing values, I decided to calculate basic statistics for numeric and categorical columns and also created a pair plot to assess whether there seem to be already clear relationships emerging.

I also decided to separate the data by whether the person is a smoker, because I am expecting that this makes a difference

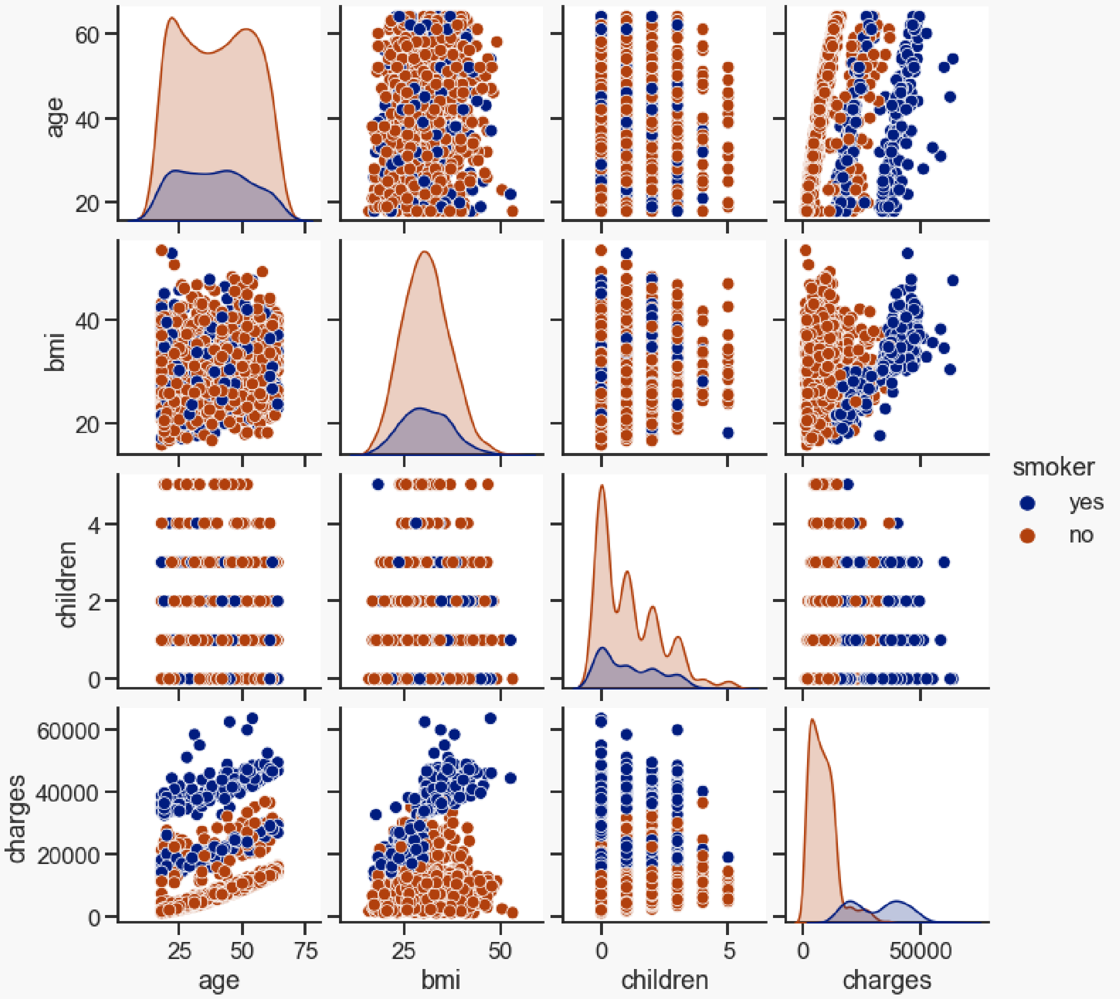


__Figure 1:__ Pairwise scatterplot of insurance dataset separated by smoker and non-smoker

From Figure 1 can be seen that some columns, such as the number of children and the insurance charges are skewed and that there indeed seems to be a differing behavior in charges between smokers and non-smokers with smokers producing on average larger charges.

The skewness was calculated as

Column Name	Skewness
charges	1.51588
children	0.93838

Further investigation shows that the dataset is approximately balanced with respect to region and sex, but that there are many more non-smokers (n=1064) than smokers (n=274) in the dataset.

Preprocessing and Feature Engineering

The charges column was designated as the target and dropped from the feature list.

Based on the structure of the dataset, I decided to:

- Perform a log transform on the *charges* and *children* columns to reduce skewness
- Create polynomial features (for second degree polynomial) for numeric columns
- Perform one-hot encoding for ordinal columns

This was executed `transformer` objects for preprocessing in sklearn:

```
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.linear_model import LinearRegression

n_deg = 2 # degree of polynomial transformation for numeric features

numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('poly', PolynomialFeatures(degree=n_deg)),
    ('scaler', StandardScaler())])

categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='constant', fill_value='missing')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)])

clf = Pipeline(steps=[('preprocessor', preprocessor),
                      ('regressor',  LinearRegression())])
```

Description of Models for Linear Regression

Subsequently 3 linear regression models were fitted to the data using kfold crossvalidation (kf-value = 3) using the same kf split for each training pipeline.

The models were:

1. Linear regression
2. Ridge regression
3. Lasso Regression

Linearization parameters (alpha) for ridge and lasso regression were estimated using grid-search.

The following R2-scores were achieved:

Model	R^2**-score**
Linear Regression	0.7772
Ridge Regression (alpha= 1.858)	0.7771
Lasso Regression (alpha = 0.000574)	0.7769

Model Selection

It is apparent from the above R^2-scores that all models have very similar performance.

Regarding explainability of results, there is also little difference.
All models have regression coefficients with absolute values between 0 and 1.6 such that there is little evidence of overfitting even for the basic linear regression.

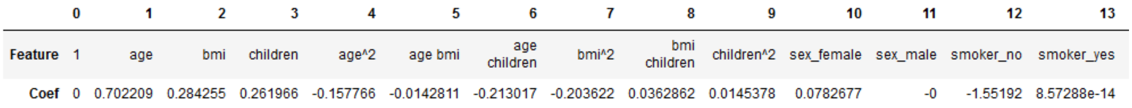
Looking at the regression coefficients for Lasso-regression it becomes apparent that the Lasso-regression eliminated the redundant information contained in the one-hot encoded features for *sex* and *smoker* to zero, by setting one of the one-hot encoded feature coefficient to zero. Other than that coefficients remained very similar between the 3 regression models.

- A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability.

Key Findings

As expected smoking and age were found to be the most important predictors for health insurance costs, with BMI and the number of children also having a positive effect on health insurance charges. Geographical region (not shown), sex, as well as polynomial features were found to have only small effects.

Table: Regression Coefficients for Lasso Regression



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Feature	1	age	bmi	children	age^2	age bmi	age children	bmi^2	bmi children	children^2	sex_female	sex_male	smoker_no	smoker_yes
Coef	0	0.702209	0.284255	0.261966	-0.157766	-0.0142811	-0.213017	-0.203622	0.0362862	0.0145378	0.0782677	-0	-1.55192	8.57288e-14

Looking at predicted versus actual charges (Figures 2 and 3 with results from Lasso-regression shown as example), it becomes apparent that all three linear models have considerable shortcomings. The model shows good performance up to charges of approximately $17,000. At higher charges the model shows considerable bias. This is considerably worse for data from smokers, for which the model is not able to capture charges very well. It overstimates insurance charges for smokers in the majority of cases. **This may be due to the fact that the training data is not-balanced with respect to smokers and that some smokers have very high actual charges, which may be considered outliers and may impact model performance.**

Conversely, for charges above $17,000 the model underpredicts charges for non-smokers.

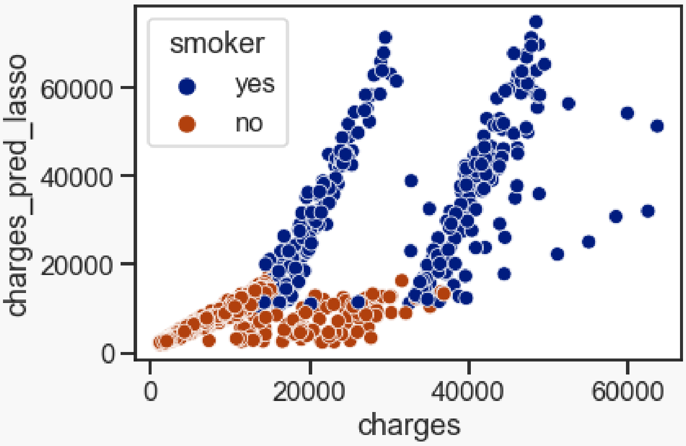


Figure 2: *Scatterplot of actual vs. predicted charges from Lasso-regression separated by smokers and non smokers*

Looking at the influence of age on predicted charges (Figure 3), it appears as if the there is a subgroup of data for which the model overestimates the impact of age on predicted charges. For **non-smokers with high charges and low age the model underpredicts charges, while it overpredicts charges for smok model is not capable for distinguishing between smokers with high charges and smokers with low charges** (the two groups apparent in Figure 2).

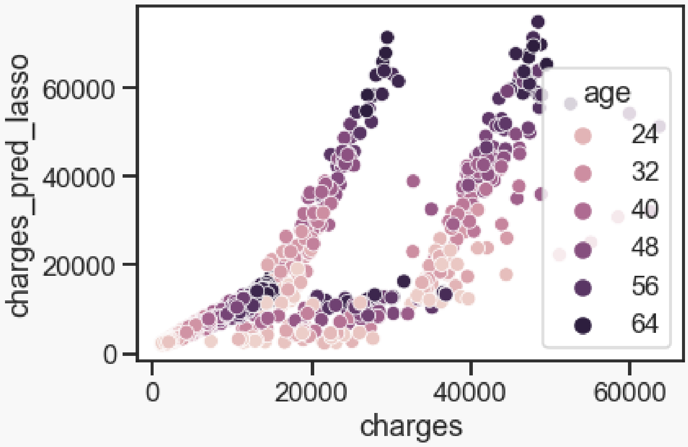


Figure 3: *Scatterplot of actual vs. predicted charges from Lasso-regression separated by age groups*

Outlook

The analysis has discovered several shortcomings of the model and specifically that the model does not capture well the behavior of the smoker group, which tends to have higher charges than the non-smokers. Given the fact that the dataset only contains 6 features, important variables that impact healthcare costs may be missing. Examples could include other risk factors like occupation, chronic disease and family history for specific health conditions.

There is some indication in the results, that BMI may be responsible (Figure not shown) with higher BMI smokers producing higher charges. At present the models lack an interaction term between the binary smoker label and BMI. Adding such an interaction term may improve model performance.

Additionally, given the fact that smokers are under-represented in the dataset, it may make sense to gather additional data from smokers or to ensure a balanced training and test datasets to improve the model for smokers. **Alternatively, it may make sense to train separate models for smokers and non smokers**

There is little evidence that the additional feature engineering of polynomical features improved model performance.

Given the bimodal structure of health insurance charges (seen in Figure 1), it is possible that linear models may not be able to capture this behavior such that non-linear methods may be needed, if the above mentioned methods do not lead to success.